Document classification/TF-IDF

Yuanjing Zhu

October 19, 2022

1 Introduction

In this assignment, I used term-document matrices and weightings along with the K-Nearest-Neighbors (KNN) algorithm to distinguish between documents from 2 categories (hobbies, romance) in the Brown corpus.

2 Result

Here are the accuracy of the document classifier for all three types of document vectors:

Document vector type	Accuracy
Raw counts	78.46%
TF-IDF weighting	78.46%
TF-IDF variant	83.08%

- **Raw counts**

  Raw counts simply count the frequency of a word appearing in the document. Although it can be helpful to understand the type of text, it fails to distinguish more important words and less important words for analysis. Most common terms like "to", "and" appear frequently, but they provide little context of the document, thus providing little help to differentiate documents. Therefore, only 78.46% is correct using raw counts as term-document vectors.

- **TF-IDF**

  The formula of TF-IDf I use is[1]:

  $$log_{10}(count(t,d)+1) * log_{10}(\frac{N}{df_t})$$

  TF-IDF is based on the idea that words that are overly common used in a corpus are not statistically significant for identifying patterns. A higher TF-IDF value denotes a word's greater importance within the corpus, whereas a lower value denotes a word's lesser importance. It is very useful for discriminating documents from the rest of the collection.

  However, the accuracy of using TF-IDF weighting (78.46%) does not increase the percentage of correctness compared with using raw content because the terms we chose are "to" and "could", which are trivial and carry little useful information about the documents.

- **TF-IDF variant**

  In this new variant of TF-IDF weighting, the formula I use is[2]:

  $$a + (1-a) * \frac{tf_{t,d}}{tf_{max}(d)} * log_{10}(\frac{N}{df_t})$$

The term a is a smoothing term, which is generally set to be 0.4. In this implementation, IDF stays the same while TF is normalized by the maximum term frequency in this document. It can deal with the issue where a long document has higher TF value simply because it repeats the same term again and again due to its length. But it can fail under the circumstance when there is outlier term which occurs exceptionally high in a document but is not typical to represent its content.

In this assignment, the length of each document is different. Among the 65 documents in Brown corpus, the maximum document length is 2546 while the minimum is 2187. Therefor, by normalizing tf by maximum term frequency, it can improve accuracy to 83.08%

3 Reference

[1] Dan Jurafsky and James H. Martin. (2021). Speech and Language Processing. Stanford University.

[2] Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to information retrieval. Cambridge university press.